IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Anh Vu **NGUYEN**
08/12/2021

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

- **Summary of methodologies:** With collected relevant data of SpaceX's rocket launches via SpaceX API and public websites, valuable insights about the first stage of Falcon 9 can be extracted. Because the data coming from multiple sources, data wrangling and data cleaning were conducted before applying exploratory data analysis (EDA). With EDA, significant correlations between different features were detected. These factors are highly important for the success of the launch. To find the relationship between these features, SQL queries and data visualization were conducted intensively. Then, interactive dashboards were created to give us a more intuitive perspective about each launch. We selected appropriate features to build the dataset for our machine learning (ML) models. During the modeling stage, multiple algorithms were examined carefully to find the most appropriate model for our ultimate objective – the prediction of Falcon 9's first stage success.
- **Summary of all results:** Launch sites and payload mass are detected key features that decide the success of Falcon 9's first stage success. Decision Tree was the ML model that yielded the best prediction accuracy. Through analysis, we can obviously recognize that SpaceX constantly improve Falcon 9 to further reduce the flight cost and increase the success rate.

Introduction

- SpaceX is currently the game changer of the aerospace manufacturing and space transportation industries with the Falcon 9. The flight cost of each launch is reduced significantly because the reusability of Falcon 9's first stage. To compete with SpaceX, predictive model for Falcon 9's first stage outcome is needed. Moreover, key features that are related to success landing need to be found.
- In this project, we aim to build a robust predictive ML model and detect key features that decide the outcome of the Falcon 9's first stage.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX REST API and web scrapping
- Perform data wrangling
  - Finding patterns and deciding target label
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- In this stage, data are collected mainly by using SpaceX REST API

  https://api.spacexdata.com/v4/

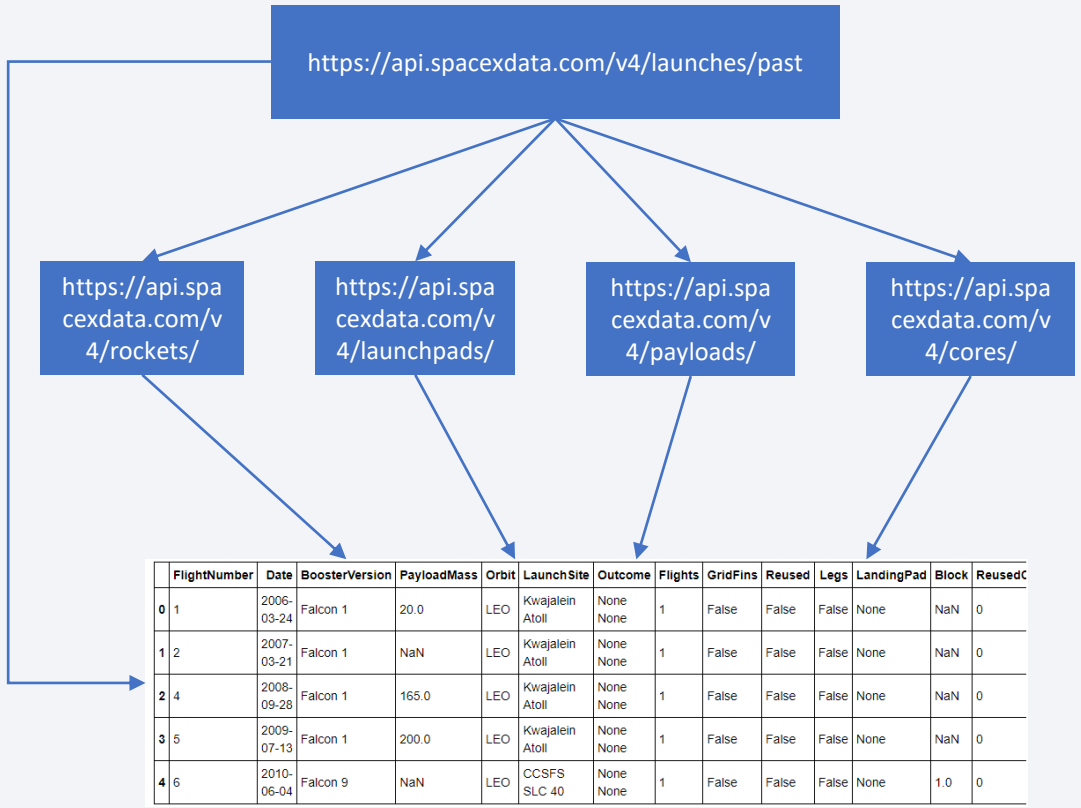
- Three endpoints that we used:
  - Rockets: detailed info about rocket versions
  - Cores: detailed info for serialized first stage cores
  - Pasts: detailed info about past launches
  - LandPads: detailed info about landing pads and ships
- More info: https://github.com/r-spacex/SpaceX-API/blob/master/docs/README.md

Data Collection – SpaceX API

https://github.com/anh-vunguyen/ibm_applied_datascience_capstone/blob/master/SpaceX_Data_Collection_API.ipynb

Functions to call endpoints and extract info:

- getBoosterVersion(): get info about Booster version
- getLaunchSite(): get info about LaunchSite
- getPayloadData(): get info about Payload
- getCoreData(): get info about Cores



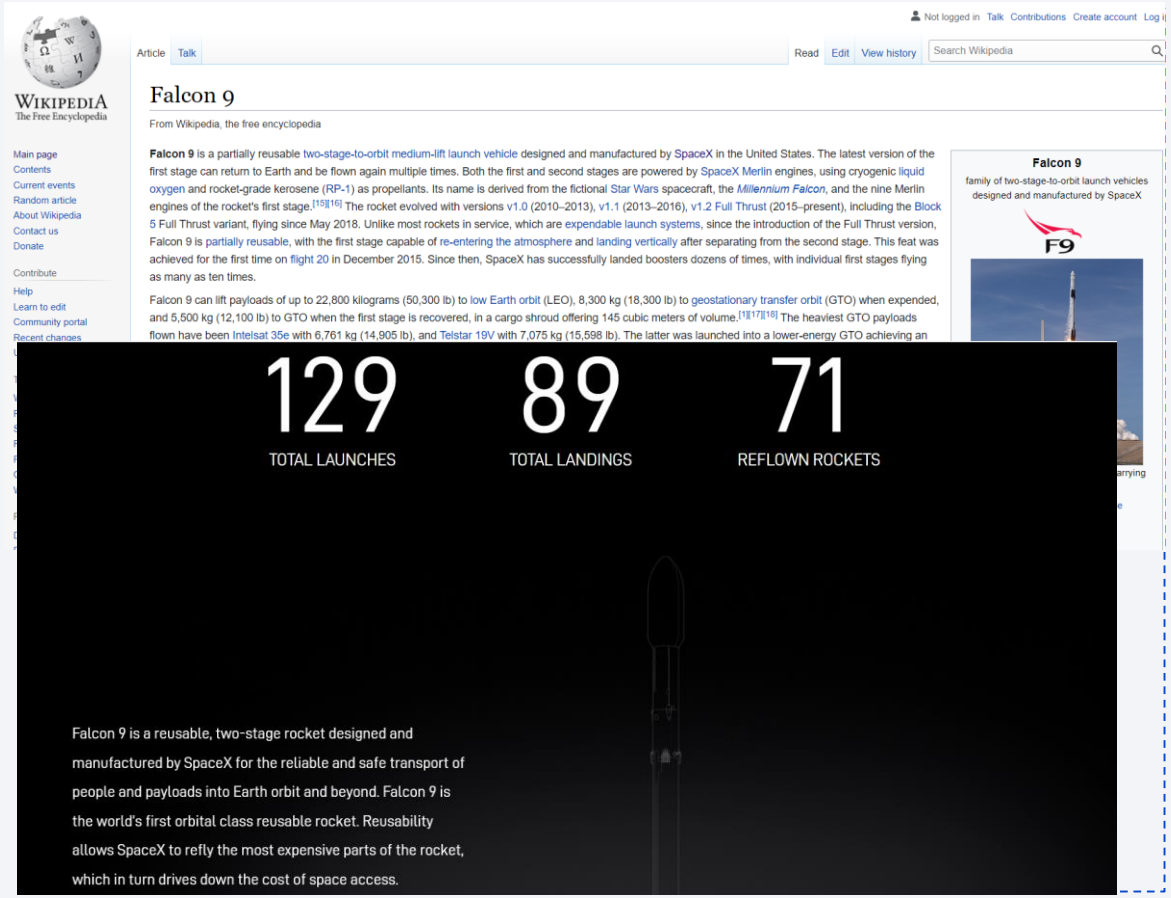
	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedC
0	1	2006-03-24	Falcon 1	20.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0
1	2	2007-03-21	Falcon 1	NaN	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0
2	4	2008-09-28	Falcon 1	165.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0
3	5	2009-07-13	Falcon 1	200.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0
4	6	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0

Data Collection - Scraping

- Because the data collected from SpaceX API are relatively complete, web scrapping is not really necessary for this project.
- Some useful websites:

  https://www.spacex.com/vehicles/falcon-9/

  https://en.wikipedia.org/wiki/Falcon_9



https://github.com/anh-vunguyen/ibm_applied_datascience_capstone/blob/master/SpaceX_Data_Collection_API.ipynb

Data Wrangling

- In this stage:
  - Identified and handled missing values
  - Calculated the number of launches on each site
  - Calculated the number and occurrence of each orbit
  - Calculated the number and occurrence of mission outcome per orbit type
  - Selected target label and created a landing outcome label from Outcome column

- https://github.com/anh-vunguyen/ibm_applied_datascience_capstone/blob/master/Data_Wrangling.ipynb

EDA with Data Visualization

- Scatter plots are utilized to show:
  - Relationship between Flight Number and Launch Site
  - Relationship between Payload and Launch Site
  - Relationship between Flight Number and Orbit type
  - Relationship between Payload and Orbit type
- Bar chart is utilized to show:
  - Relationship between success rate of each orbit type
- Line plot is utilized to show:
  - Launch success yearly trend
- https://github.com/anh-vunguyen/ibm_applied_datascience_capstone/blob/master/Exploratory_Analysis_with_Visualization.ipynb

EDA with SQL

- In this stage, SQL queries were used to help us understand the data
  - Display the names of the unique launch sites
  - Display 5 records where launch sites begin with the string 'CCA‘
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display average payload mass carried by booster version F9 v1.1
  - List the date when the first successful landing outcome in ground pad
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - List the total number of successful and failure mission outcomes
  - List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

- https://github.com/anh-vunguyen/ibm_applied_datascience_capstone/blob/master/Exploratory_Data_Analysis.ipynb

Build an Interactive Map with Folium

- By using Folium, we created an interactive Map to analyze geospatial information about the data. Moreover, some calculations were conducted to help us understand better about the launch sites.
  - Created and added folium.Circle and folium.Marker for each launch site on the site map
  - Marked the success/failed launches for each site on the map with MarkerCluster
  - Calculated the distances between a launch site to its proximities

- https://github.com/anh-vunguyen/ibm_applied_datascience_capstone/blob/master/Visualization_with_Folium.ipynb

Build a Dashboard with Plotly Dash

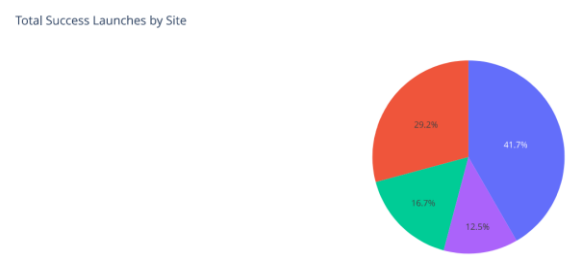
- Dashboard is an extremely useful tool for monitoring the streaming data. Besides, it is also an interactive tool to show static data for our project.
- Plotly Dash was employed to create a web-based dashboard.
  - Dropdown list to enable Launch Site selection
  - Pie chart to show the total successful launches count for all sites
  - Slider to select Payload range
  - Scatter plot to show the correlation between payload and launch success
- https://github.com/anh-vunguyen/ibm_applied_datascience_capstone/blob/master/spacex_dash_app.py

Predictive Analysis (Classification)

- After loading the dataset, categorical features were handled by using One-Hot encoding. We split the prepared dataset into train set and test set (80% / 20%).
- Multiple ML algorithms were examined:
  - Logistic Regression
  - Support Vector Machine
  - Decision Tree
  - K-Nearest Neighbors
- Grid Search strategy was employed to find optimal parameters for the models.
- Some stronger algorithms can be employed in this step: XGBoost, Random Forest and Multi-Layer Perceptron Neural Network.
- https://github.com/anh-vunguyen/ibm_applied_datascience_capstone/blob/master/Machine_Learning_Prediction.ipynb

Results

- Launch site and Payload mass are detected key features that decide the success of Falcon 9's first stage success. Decision Tree was the ML model that yielded the best prediction accuracy. Through analysis, we can obviously recognize that SpaceX constantly improve Falcon 9 to further reduce the flight cost and increase the success rate.

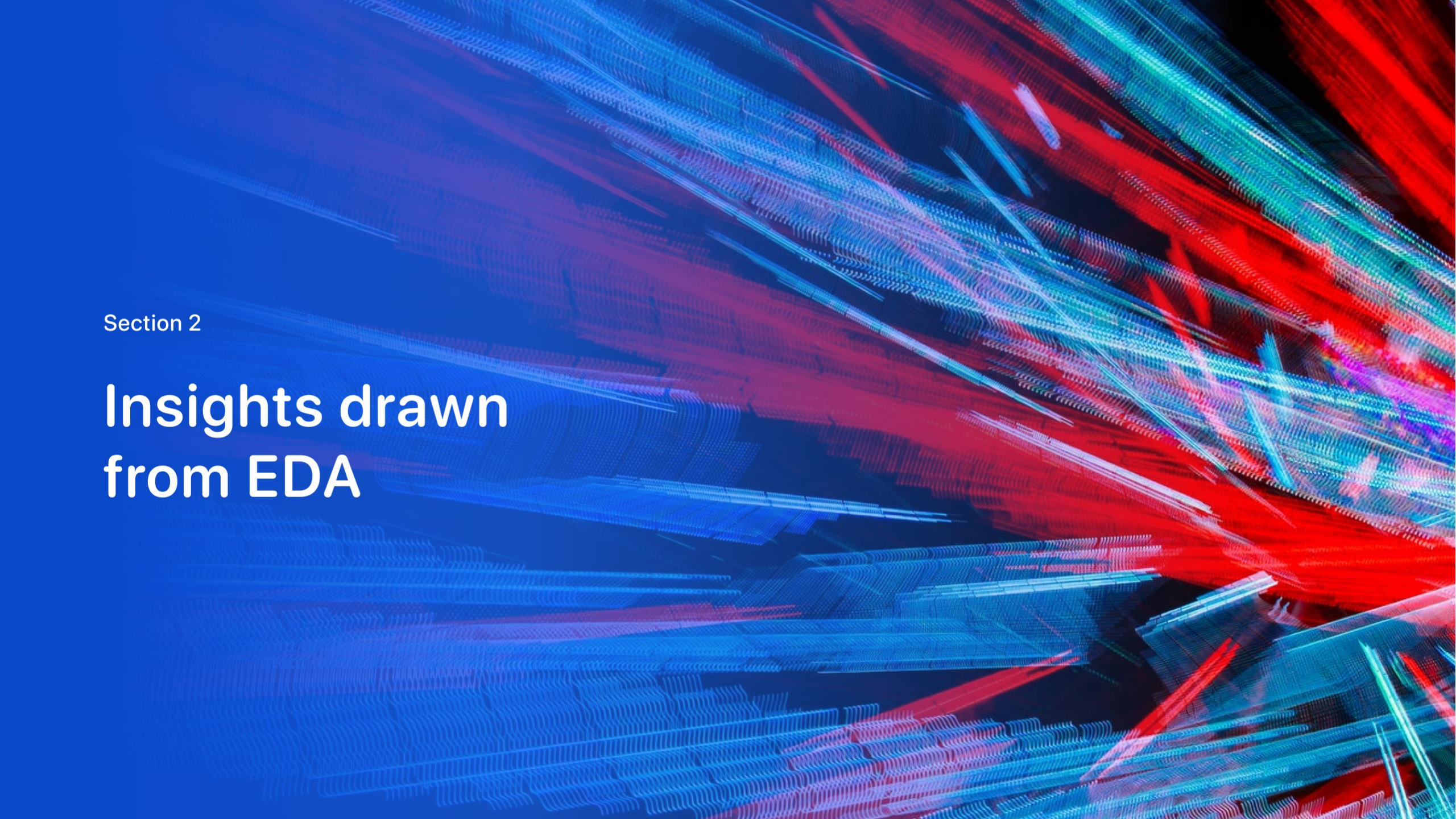


- Predictive analysis results: Decision Tree model can yield predictive landing outcome up to 88% accuracy.

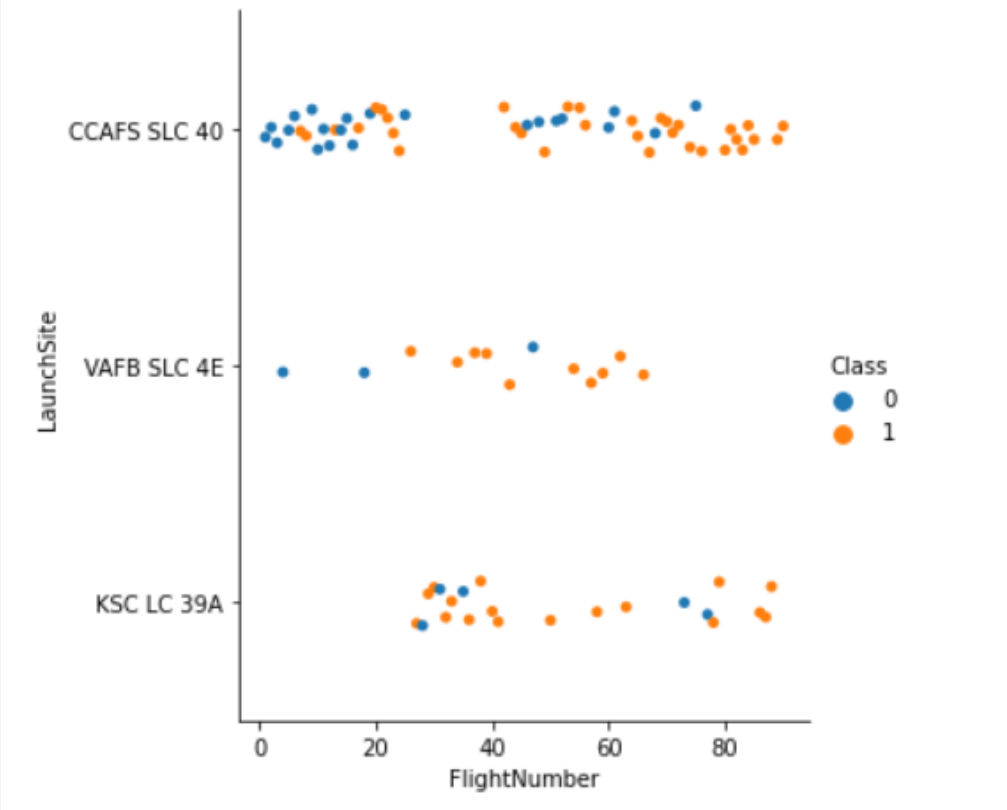
```
In [19]:  tree_score = tree_cv.score(X_test, Y_test)
          tree_score

Out[19]:  0.8888888888888888
```

Section 2

Insights drawn from EDA

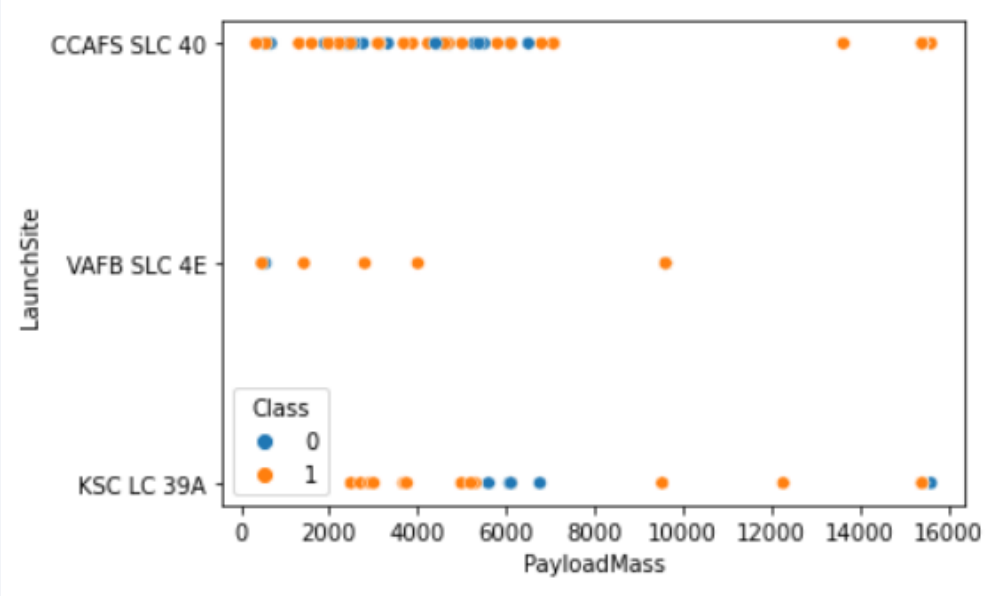


Flight Number vs. Launch Site



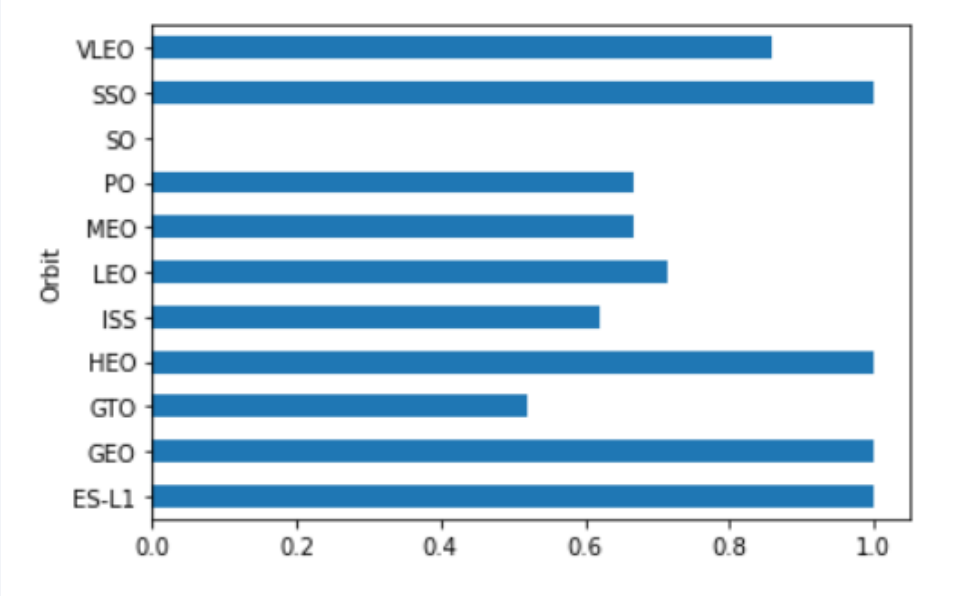
- Recent flights are more likely to be successful.
- A large number of flights were operated in CCAFS SLC 40. For recent flights, CCAFS SLC 40 and KSC LC 39A were heavily used.

Payload vs. Launch Site



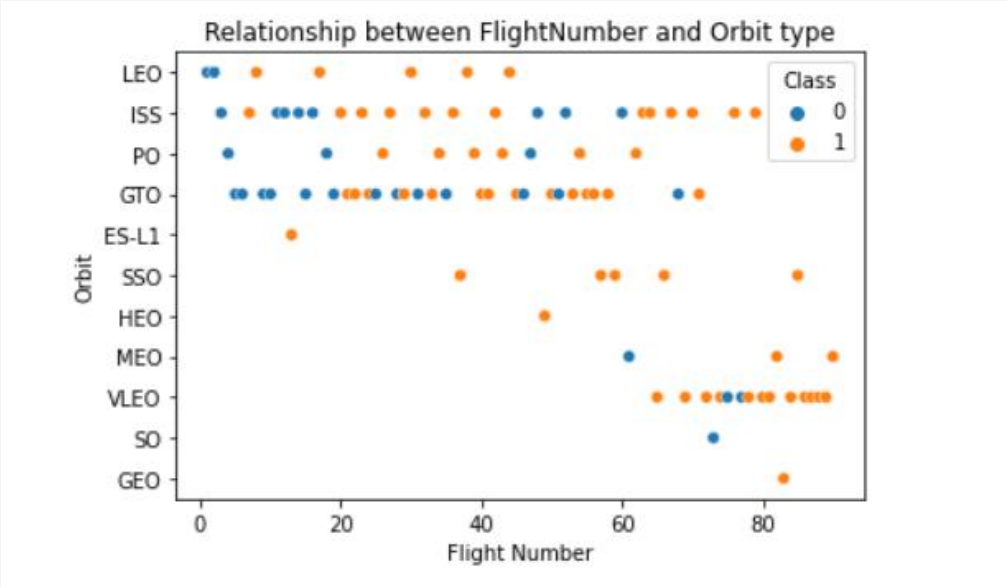
- CCAFS SLC 40 and KSC LC 39A were used for all range of payload. Whereas, VAFB SLC 4E was used for small to medium payload.

Success Rate vs. Orbit Type



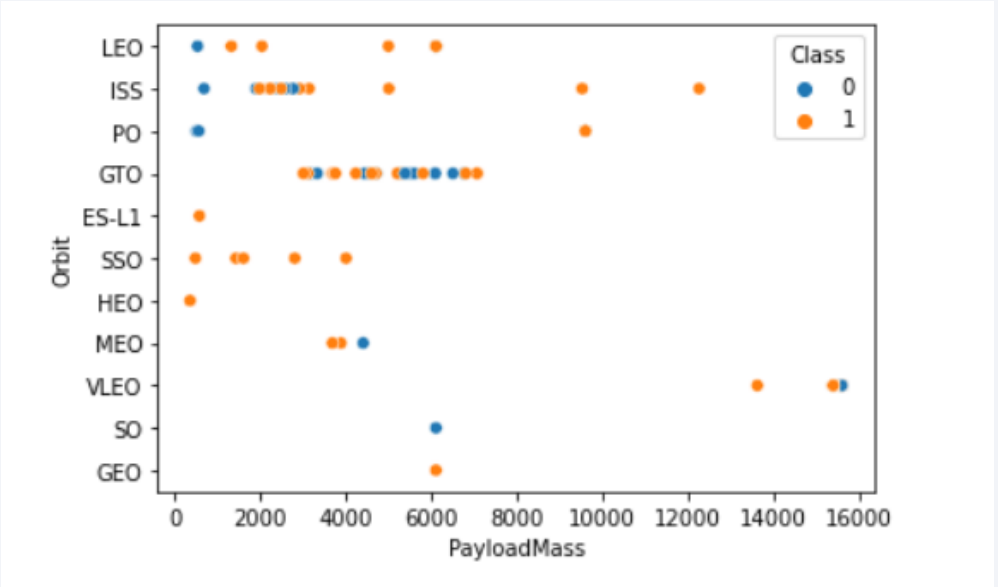
- High success rates were achieved with orbits such as VLEO, SSO, HEO, GEO, and ES-L1.
- Relatively low success rates were achieved in PO, MEO, LEO, ISS, and GTO.
- SpaceX is still not successful at SO.

Flight Number vs. Orbit Type



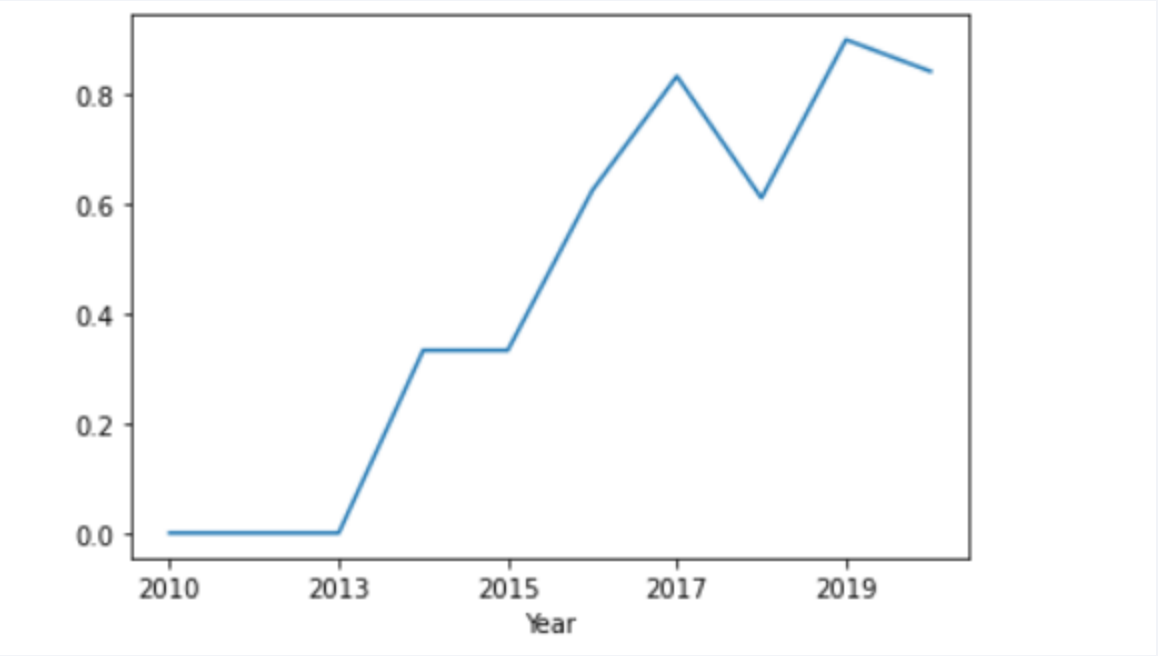
- SpaceX is currently successful at all types of orbit except SO.

Payload vs. Orbit Type



- For heavy payloads, the successful landing or positive landing rate are more likely for Polar, LEO and ISS.
- Uncertainty can be found in GTO.

Launch Success Yearly Trend



- Since 2013, SpaceX tends to become better and better over years.
- Uncertainty can be found in GTO.

All Launch Site Names

Display the names of the unique launch sites in the space mission

In [5]:
```
%%sql
SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL;
```

```
 * ibm_db_sa://bmz67690:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

Out[5]:

launch_site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

- CCAFS LC-40: Cape Canaveral Space Launch Complex 40, Florida, USA
- CCAFS SLC-40 (previously LC-40): Cape Canaveral Space Launch Complex 40, Florida, USA
- KSC LC-39A: Kennedy Space Center Launch Complex 39, Florida, USA
- VAFB SLC-4E: Vandenberg Space Launch Complex 4, California, USA

Launch Site Names Begin with 'CCA'

Display 5 records where launch sites begin with the string 'CCA'

In [6]:
```
%%sql
SELECT * FROM SPACEXTBL
WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5;
```

```
 * ibm_db_sa://bmz67690:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

Out[6]:

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

Display the total payload mass carried by boosters launched by NASA (CRS)

In [8]:
```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL
GROUP BY CUSTOMER
HAVING CUSTOMER = 'NASA (CRS)';
```

```
 * ibm_db_sa://bmz67690:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

Out[8]:

1
45596

Average Payload Mass by F9 v1.1

Display average payload mass carried by booster version F9 v1.1

In [10]:
```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) FROM
(SELECT PAYLOAD_MASS__KG_ FROM SPACEXTBL
WHERE booster_version LIKE 'F9 v1.1%');
```

```
 * ibm_db_sa://bmz67690:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

Out[10]:

1
2534

- The average payload carried by F9 v1.1 is 2534 kg.

First Successful Ground Landing Date

List the date when the first successful landing outcome in ground pad was acheived.

Hint:Use min function

In [12]:
```
%%sql
SELECT MIN(DATE) FROM SPACEXTBL
WHERE LANDING__OUTCOME = 'Success (ground pad)';
```

```
 * ibm_db_sa://bmz67690:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

Out[12]:

1
2015-12-22

- 22 Dec 2015 is truly a historical milestone not only for SpaceX but also for the aerospace transportation industry.

Successful Drone Ship Landing with Payload between 4000 and 6000

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

In [16]:
```
%%sql
SELECT booster_version FROM SPACEXTBL
WHERE LANDING__OUTCOME = 'Success (drone ship)' AND (payload_mass__kg_ > 4000 AND payload_mass__kg_ < 6000);
```

```
 * ibm_db_sa://bmz67690:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

Out[16]:

booster_version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

In [18]:
```
%%sql
SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) FROM SPACEXTBL
GROUP BY MISSION_OUTCOME;
```

```
 * ibm_db_sa://bmz67690:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

Out[18]:

mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

In [21]:
```
%%sql
SELECT booster_version, payload_mass__kg_ FROM SPACEXTBL
WHERE payload_mass__kg_ = (SELECT MAX(payload_mass__kg_) FROM SPACEXTBL);
```

```
 * ibm_db_sa://bmz67690:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

Out[21]:

booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

In [23]:
```
%%sql
SELECT landing__outcome, booster_version, launch_site FROM SPACEXTBL
WHERE YEAR(DATE) = 2015;
```

```
 * ibm_db_sa://bmz67690:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

Out[23]:

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Controlled (ocean)	F9 v1.1 B1013	CCAFS LC-40
No attempt	F9 v1.1 B1014	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40
No attempt	F9 v1.1 B1016	CCAFS LC-40
Precluded (drone ship)	F9 v1.1 B1018	CCAFS LC-40
Success (ground pad)	F9 FT B1019	CCAFS LC-40

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

In [28]:

```
%%sql
SELECT landing__outcome, COUNT(landing__outcome) FROM
(SELECT * FROM SPACEXTBL
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20')
GROUP BY landing__outcome
ORDER BY COUNT(landing__outcome) DESC;
```

```
 * ibm_db_sa://bmz67690:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

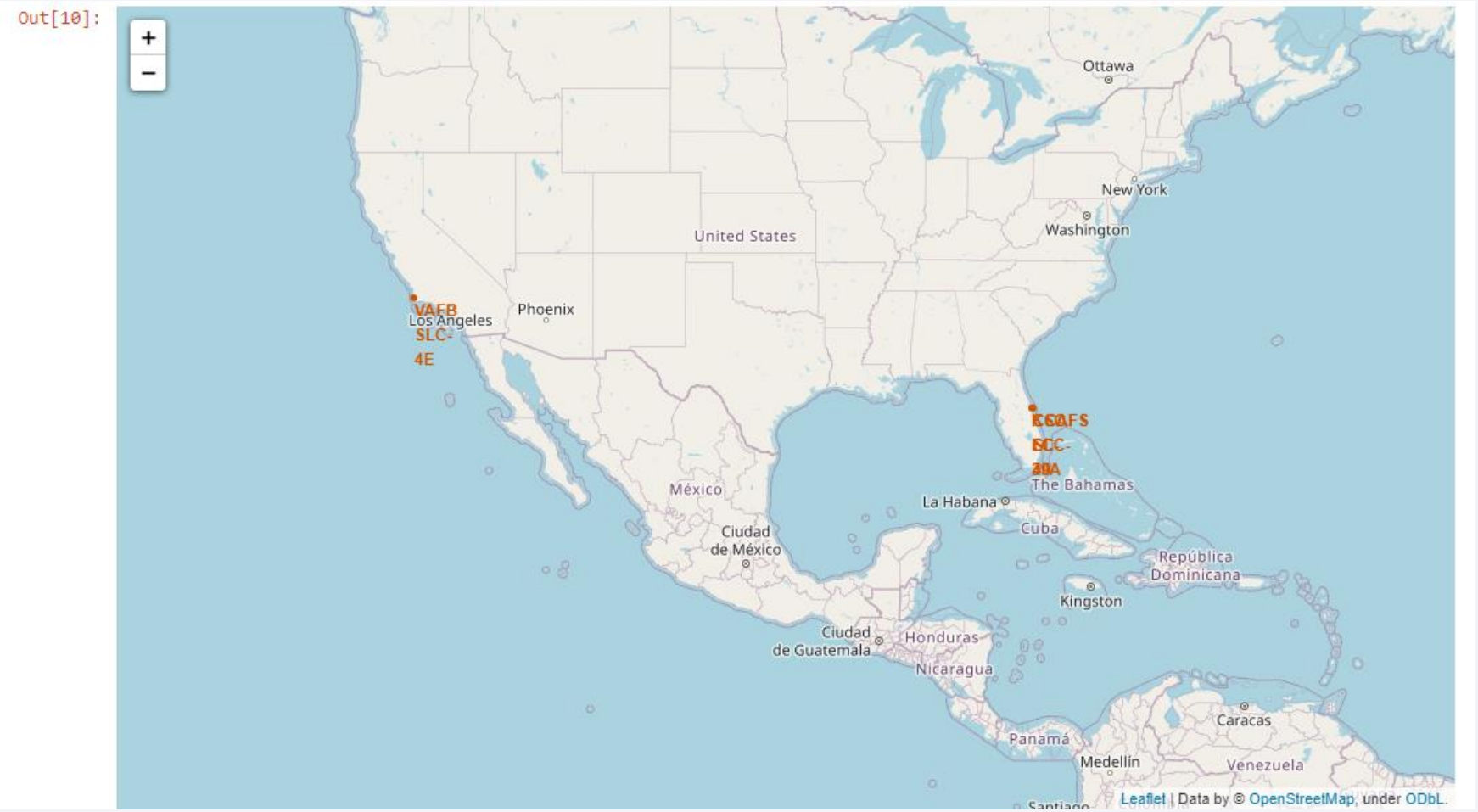
Out[28]:

landing__outcome	2
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 4

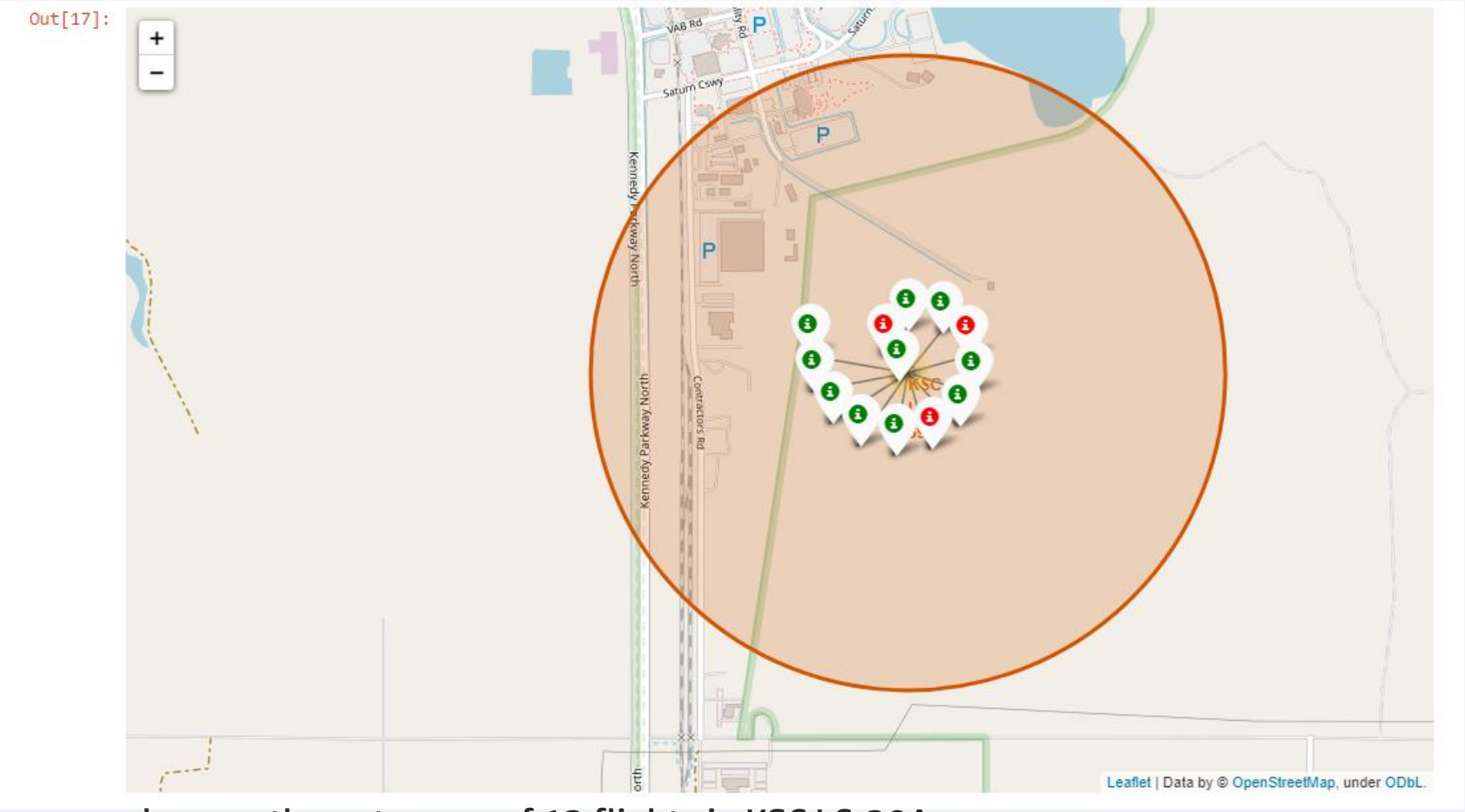
Launch Sites Proximities Analysis

<Folium Map Screenshot 1>



Only VAFB is in California while the others are in Florida. By zooming in, we can find out that CCAFS LC-40 and CCAFS SLC-40 are in the same location.

<Folium Map Screenshot 2>



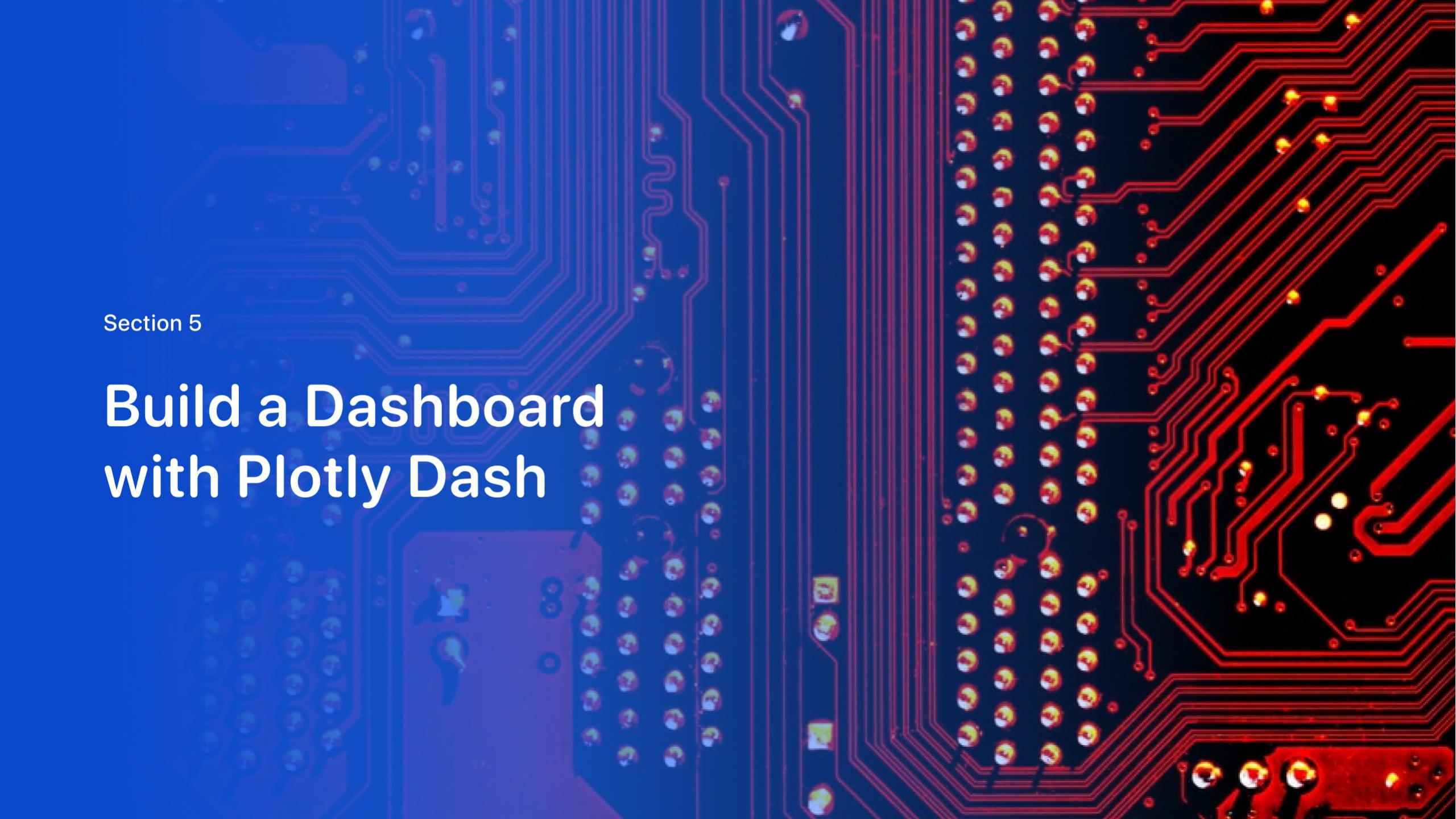
For example, we can observe the outcomes of 13 flights in KSC LC-39A

<Folium Map Screenshot 3>

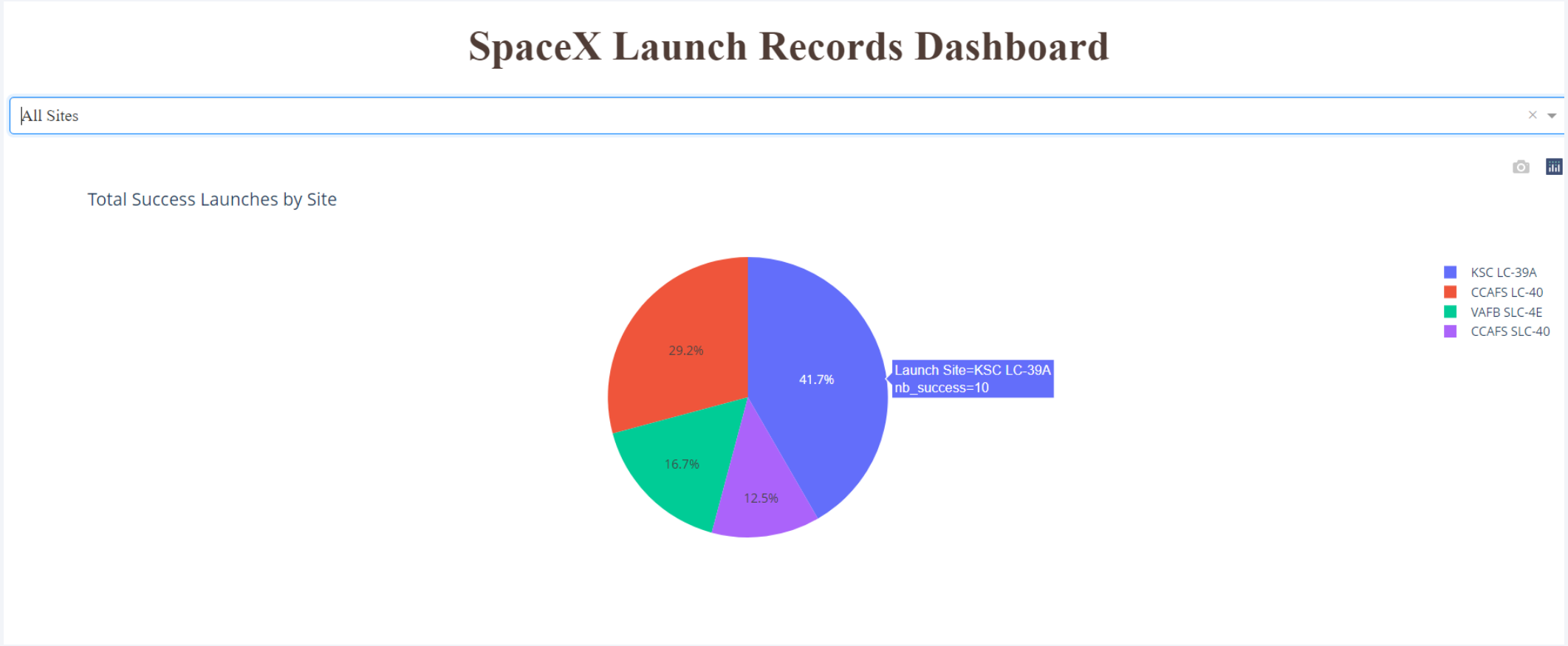
The distance from CCAFS SLC-40 to the nearest coastline is nearly 0.9 Km.

Section 5

Build a Dashboard with Plotly Dash

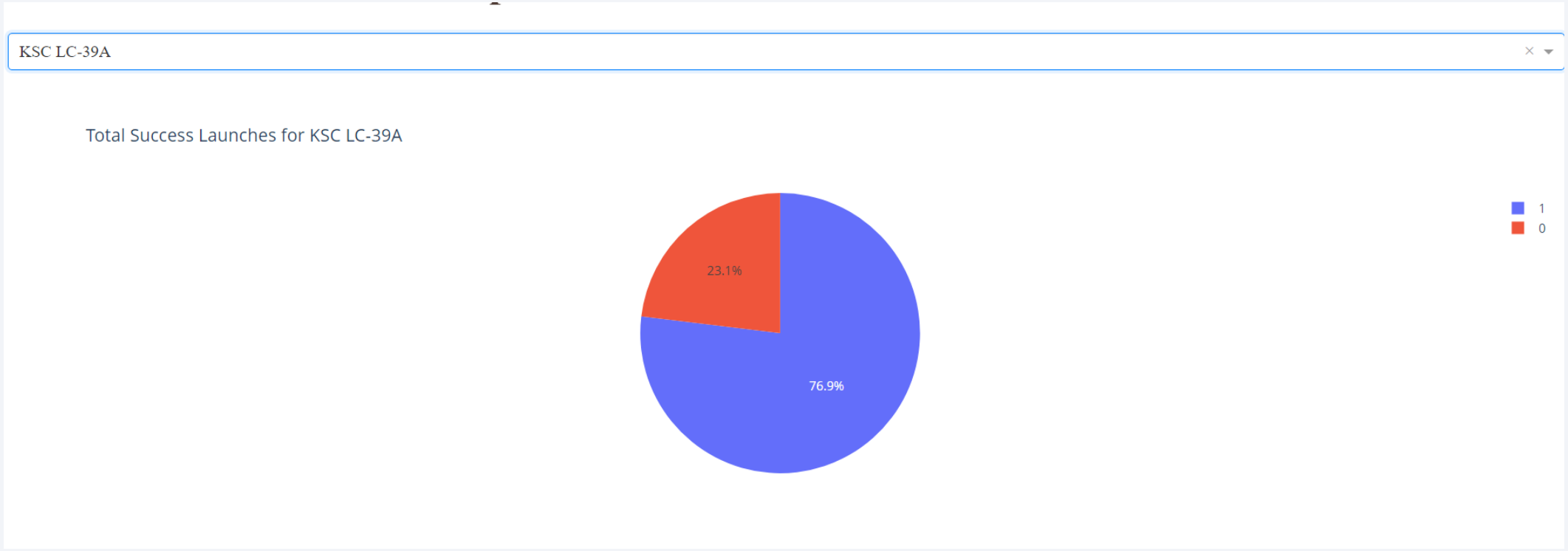


<Dashboard Screenshot 1>



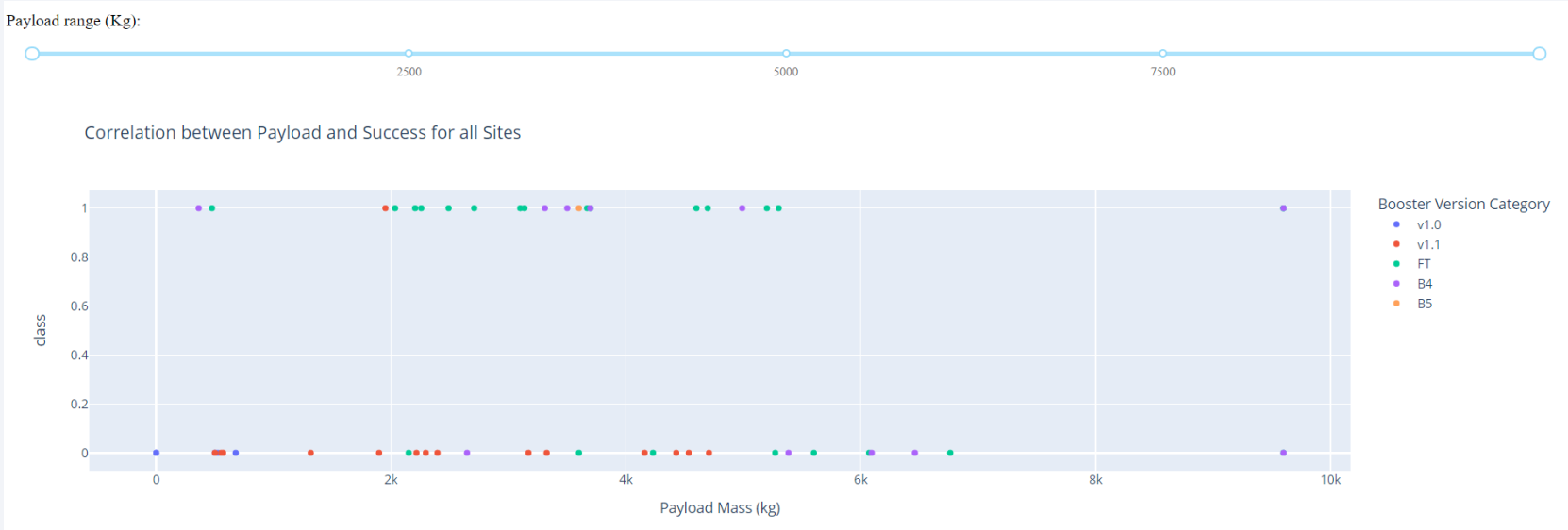
KSC LC-39A is the launch site that has the highest number of success flights.

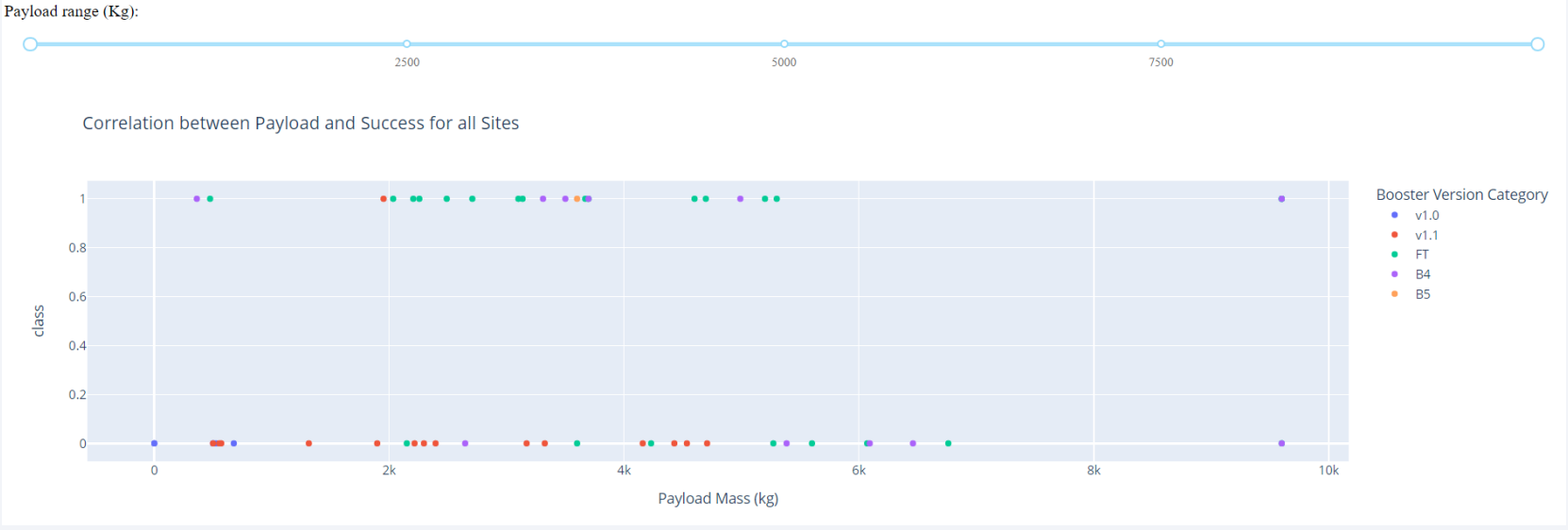
<Dashboard Screenshot 2>



In KSC LC-39A, the success rate is up to 77%!

<Dashboard Screenshot 3>





Section 6

Predictive Analysis (Classification)

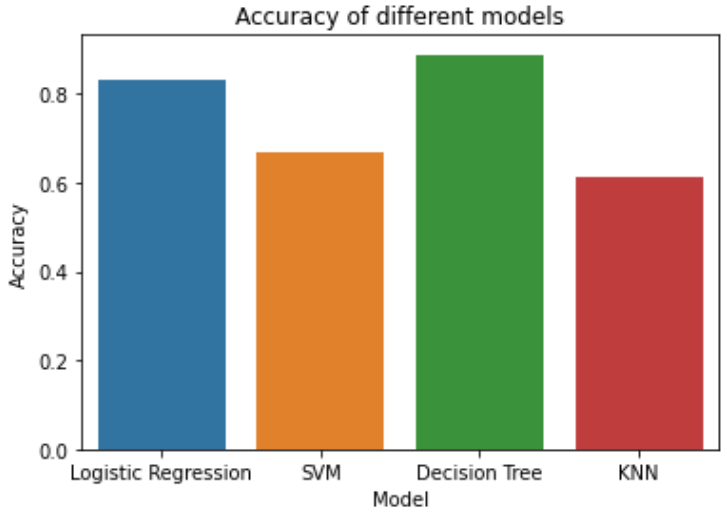
Classification Accuracy

In [39]:
```
scores = {'Logistic Regression': logreg_cv_score, 'SVM': svm_score, 'Decision Tree': tree_score, 'KNN': KNN_score}
print("The best model is: ", max(scores, key=scores.get), "with the accuracy of ", max(scores.values())*100, "%")
```

```
The best model is:  Decision Tree with the accuracy of  88.88888888888889 %
```

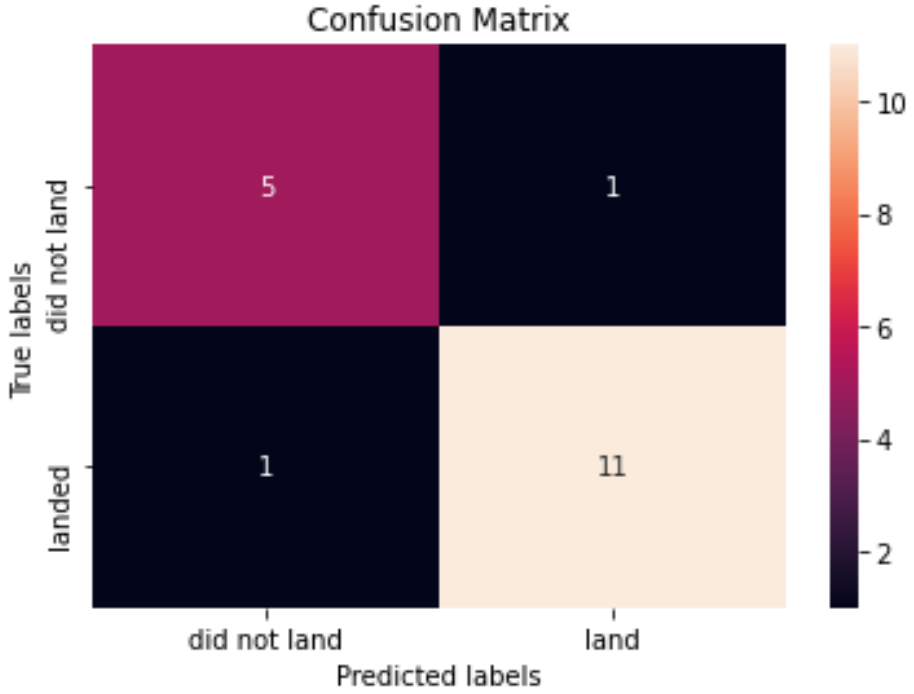
In [37]:
```
sns.barplot(x=list(scores.keys()), y=list(scores.values()))
plt.title("Accuracy of different models")
plt.xlabel("Model")
plt.ylabel("Accuracy")
```

Out[37]: Text(0, 0.5, 'Accuracy')



Confusion Matrix

```
In [25]: yhat = tree_cv.predict(X_test)
         plot_confusion_matrix(Y_test,yhat)
```



Decision Tree model can predict up to 88.8% accuracy.

Based on the confusion matrix, our model predict correctly 11 “land” outcomes and 5 “did not land” outcomes. Only 2 flights were incorrect.

Conclusions

- Launch sites and payload mass are detected key features that decide the success of Falcon 9's first stage success.
- KSC LC-39A - Kennedy Space Center Launch Complex 39 is the launch site with highest success rate.
- Decision Tree was the ML model that yielded the best prediction accuracy.
- Through analysis, we can obviously recognize that SpaceX constantly improve Falcon 9 to further reduce the flight cost and increase the success rate.

Appendix

- As we know, weather condition is extremely important during rocket launching. In the future project, data of weather station in the launch site are recommended to add in order to build much better model. Besides, we can analyze which are the ideal weather (temperature, humidity, wind speed, etc) for launching rocket.
- Some stronger algorithms can be added to the future projects such as XGBoost, Random Forest, etc.

Thank you!